High Fidelity 3D Hand Shape Reconstruction via Scalable Graph Frequency Decomposition

Tianyu Luan[1] Yuanhao Zhai[1] Jingjing Meng[1] Zhong Li[2]
Zhang Chen[2] Yi Xu[2] Junsong Yuan[1]

[1]State University of New York at Buffalo [2]OPPO US Research Center, InnoPeak Technology, Inc.
{tianyulu,yzhai6,jmeng2,jsyuan}@buffalo.edu
{zhong.li,zhang.chen,yi.xu}@oppo.com

Abstract

Despite the impressive performance obtained by recent single-image hand modeling techniques, they lack the capability to capture sufficient details of the 3D hand mesh. This deficiency greatly limits their applications when high-fidelity hand modeling is required, e.g.*, personalized hand modeling. To address this problem, we design a frequency split network to generate 3D hand mesh using different frequency bands in a coarse-to-fine manner. To capture high-frequency personalized details, we transform the 3D mesh into the frequency domain, and propose a novel frequency decomposition loss to supervise each frequency component. By leveraging such a coarse-to-fine scheme, hand details that correspond to the higher frequency domain can be preserved. In addition, the proposed network is scalable, and can stop the inference at any resolution level to accommodate different hardware with varying computational powers. To quantitatively evaluate the performance of our method in terms of recovering personalized shape details, we introduce a new evaluation metric named Mean Signal-to-Noise Ratio (MSNR) to measure the signal-to-noise ratio of each mesh frequency component. Extensive experiments demonstrate that our approach generates fine-grained details for high-fidelity 3D hand reconstruction, and our evaluation metric is more effective for measuring mesh details compared with traditional metrics. The code is available at* https://github.com/tyluann/FreqHand.

1. Introduction

High-fidelity and personalized 3D hand modeling have seen great demand in 3D games, virtual reality, and the emerging Metaverse, as it brings better user experiences, *e.g.*, users can see their own realistic hands in the virtual space instead of the standard avatar hands. Therefore, it is of great importance to reconstruct high-fidelity hand meshes that can adapt to different users and application scenarios.

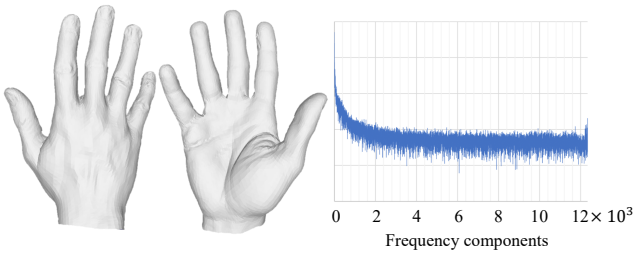

Figure 1. An exemplar hand mesh of sufficient details and its graph frequency decomposition. The x-axis shows frequency components from low to high. The y-axis shows the amplitude of each component in the logarithm. At the frequency domain, the signal amplitude generally decreases as the frequency increases.

Despite previous successes in 3D hand reconstruction and modeling [3, 6, 7, 16, 22, 40, 44, 46], few existing solutions focus on enriching the details of the reconstructed shape, and most current methods fail to generate consumer-friendly high-fidelity hands. When we treat the hand mesh as graph signals, like most natural signals, the low-frequency components have larger amplitudes than those of the high-frequency parts, which we can observe in a hand mesh spectrum curve (Fig. 1). Consequently, if we generate the mesh purely in the spatial domain, the signals of different frequencies could be biased, thus the high-frequency information can be easily overwhelmed by its low-frequency counterpart. Moreover, the wide usage of compact parametric models, such as MANO [32], has limited the expressiveness of personalized details. Even though MANO can robustly estimate the hand pose and coarse shape, it sacrifices hand details for compactness and robustness in the parameterization process, so the detail expression ability of MANO is suppressed.

To better model detailed 3D shape information, we transform the hand mesh into the graph frequency domain, and

design a frequency-based loss function to generate high-fidelity hand mesh in a scalable manner. Supervision in the frequency domain explicitly constrains the signal of a given frequency band from being influenced by other frequency bands. Therefore, the high-frequency signals of hand shape will not be suppressed by low-frequency signals despite the amplitude disadvantage. To improve the expressiveness of hand models, we design a new hand model of $12,337$ vertices that extends previous parametric models such as MANO with nonparametric representation for residual adjustments. While the nonparametric residual expresses personalized details, the parametric base ensures the overall structure of the hand mesh, *e.g*., reliable estimation of hand pose and 3D shape. Instead of fixing the hand mesh resolution, we design our network architecture in a coarse-to-fine manner with three resolution levels U-net for scalability. Different levels of image features contribute to different levels of detail. Specifically, we use low-level features in high-frequency detail generation and high-level features in low-frequency detail generation. At each resolution level, our network outputs a hand mesh with the corresponding resolution. During inference, the network outputs an increasingly higher resolution mesh with more personalized details step-by-step, while the inference process can stop at any one of the three resolution levels.

In summary, our contributions include the following.

1. We design a high-fidelity 3D hand model for reconstructing 3D hand shapes from single images. The hand representation provides detailed expression, and our frequency decomposition loss helps to capture the personalized shape information.
2. To enable computational efficiency, we propose a frequency split network architecture to generate high-fidelity hand mesh in a scalable manner with multiple levels of detail. During inference, our scalable framework supports budget-aware mesh reconstruction when the computational resources are limited.
3. We propose a new metric to evaluate 3D mesh details. It better captures the signal-to-noise ratio of all frequency bands to evaluate high-fidelity hand meshes. The effectiveness of this metric has been validated by extensive experiments.

We evaluate our method on the InterHand2.6M dataset [29]. In addition to the proposed evaluation metrics, we also evaluate mean per joint position error (MPJPE) and mesh Chamfer distance (CD). Compared to MANO and other baselines, our proposed method achieves better results using all three metrics.

2. Related Work

Parametric hand shape reconstruction. Parametric models are a popular approach in hand mesh reconstruction. Romero *et al*. [32] proposed MANO, which uses a set of shape and pose parameters to control the movement and deformation of human hands. Many recent works [16, 31, 40, 41, 44, 48–50] combined deep learning with MANO. They use features extracted from the RGB image as input, CNN to get the shape and pose parameters, and eventually these parameters to generate hand mesh. These methods make use of the strong prior knowledge provided by the hand parametric model, so that it is convenient to train the networks and the results are robust. However, the parametric method limits the mesh resolution and details of hands.

Non-parametric hand shape reconstruction. Non-parametric hand shape reconstruction typically estimates the vertex positions of a template with fixed topology. For example, Ge *et al*. [13] proposed a method using a graph convolution network. It uses a predefined upsampling operation to build a multi-level spectrum GCN network. Kulon *et al*. [21] used spatial GCN and spiral convolution operator for mesh generation. Moon *et al*. [27] proposed a pixel-based approach. However, none of these works paid close attention to detailed shapes. Moon *et al*. [28] provided an approach that outputs fine details, but since they need the 3D scanned meshes of the test cases for training, their model cannot do cross-identity reconstruction. In our paper, we design a new hand model that combines the strength of both parametric and non-parametric approaches. We use this hand model as a basis to reconstruct high-fidelity hands.

Mesh frequency analysis. Previous works mainly focused on the spectrum analysis of the entire mesh graph. Chung. [10] defines the graph Fourier transformation and graph Laplacian operator, which builds the foundation of graph spectrum analysis. [38] extends commonly used signal processing operators to graph space. [5] proposes a spectrum graph convolution network based on graph spectrum characteristics. The spectral decomposition of the graph function is used to define graph-based convolution. Recent works such as [11, 18, 24, 35, 37, 43, 51] widely use spectrum GCN in different fields. However, these works mainly focus on the analysis of the overall graph spectrum. In this paper, we use spectrum analysis as a tool to design our provided loss function and metric.

3. Proposed Method

We propose a scalable network that reconstructs the detailed hand shape, and use frequency decomposition loss to acquire details. Fig. 2 shows our network architecture. We design our network in the manner of a U-net. First, we generate a MANO mesh from image features from EfficientNet [39]. Based on the MANO mesh, we use a graph convolution network (green, yellow, and red modules in Fig. 2) to recover a high-fidelity hand mesh. In order to obtain high-frequency information, we use image features from different layers of the backbone network as a part of

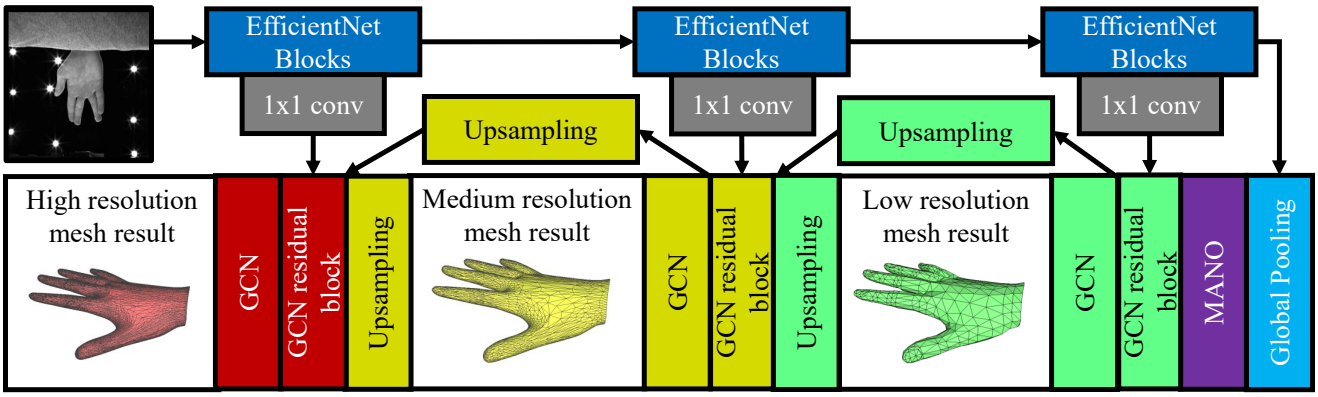

Figure 2. We design our scalable hand modeling network in a U-net manner. First, we generate a MANO mesh from image features (purple block). Then, based on the MANO mesh, we use a multilevel GCN to recover 3 levels of personalized mesh (green, yellow, and red blocks from low to high). In order to obtain high-frequency hand details, we use skip-connected image features from different layers of the backbone network (blue and gray blocks) At inference, our network can stop at any resolution level, but still provides reasonable high-fidelity results at that resolution. The architecture and implementation details can be found in supplementary material Section I and II.

the GCN inputs. Specifically, at the low-resolution level, we take high-level image features as part of the input, and use a low-resolution graph topology to generate a low-resolution mesh. At medium and high-frequency levels, we use lower-level image feature through the skip connection to produce a high-resolution mesh. Note that at every resolution level, the network will output the intermediate hand mesh, so it would naturally have the ability for scalable inference. During the training process, we supervise both intermediate meshes and the final high-resolution mesh. We discuss the details in the following.

3.1. High Fidelity 3D Hand Model

We design our hand representation based on MANO [32]. MANO factorizes human hands into a 10-dimensional shape representation β and a 35-dimensional pose representation θ. MANO model can be represented as

$$\begin{cases} M(\theta,\beta) = W(T_P(\theta,\beta),\theta,w) \\ T_P(\theta,\beta) = \overline{T} + B_S(\beta) + B_P(\theta) \end{cases} \tag{1}$$

where W is the linear blend skinning function. Model parameter w is the blend weight. B_S and B_P are another two parameters of MANO named shape blend shape and pose blend shape, which are related to pose and shape parameters, respectively. MANO can transfer complex hand surface estimation into a simple regression of a few pose and shape parameters. However, MANO has limited capability in modeling shape detail. It is not only limited by the number of pose and shape dimensions (45), but also by the number of vertices (778). In our work, we designed a new parametric-based model with 12,338 vertices generated from MANO via subdivision. The large vertex number greatly enhances the model's ability to represent details.

Subdivided MANO. To address this problem. We design an extended parametric model that can better represent details. First, we add detail residuals to MANO as

$$\begin{aligned} M'(\theta,\beta,d) &= W(T'_P(\theta,\beta,d),\theta,w'), \\ T'_P(\theta,\beta,d) &= \overline{T}' + B'_S(\beta) + B'_P(\theta) + d, \end{aligned} \tag{2}$$

where, w', $\overline{T}'$, $B'_S(\beta)$, and $B'_P(\theta)$ are the parameters our model, and d is the learnable per-vertex location perturbation. The dimension of d is the same as the number of vertices.

Besides vertex residuals, we further increase the representation capability of our hand model by increasing the resolution of the mesh. Motivated by the traditional Loop subdivision [23], we propose to design our parametric hand model by subdividing the MANO template. Loop subdivision can be represented as

$$\overline{T}' = \mathbf{L_s}\overline{T}, \tag{3}$$

where, $\overline{T}$ is original template mesh with n vertices and m edges. $\overline{T}'$ is the subdivided template mesh with $n+m$ vertices. $\mathbf{L_s} \in \mathbb{R}^{(n+m)\times m}$ is the linear transformation that defines the subdivision process. The position of each vertex on the new mesh is only determined by the neighbor vertices on the original mesh, so $\mathbf{L_s}$ is sparse. We use similar strategies to calculate B_S and B_P. The MANO parameters map the input shape and pose into vertex position adjustments. These mappings are linear matrices of dimension $x \times n$. Therefore, we can calculate the parameters as

$$\begin{aligned} w' &= (\mathbf{L_s}w^\top)^\top, \\ B'_S &= (\mathbf{L_s}B_S^\top)^\top, \\ B'_P &= (\mathbf{L_s}B_P^\top)^\top. \end{aligned} \tag{4}$$

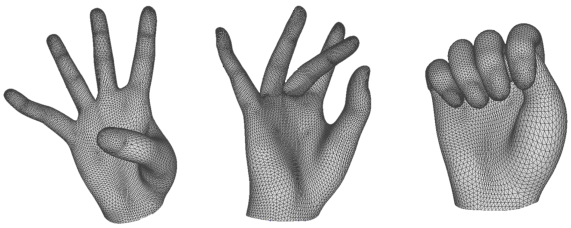

Figure 3. We design a new high-fidelity hand mesh with $12,337$ vertices. Our new model inherits the advantage of the parametric hand model and provides reliable 3D shape estimation with fewer flaws when hand poses change.

We repeat the procedure twice to get sufficient resolution.

Fig. 3 shows example meshes from the new model in different poses (d is set to 0). We can see that our representation inherits the advantages of the parametric hand model. It has a plausible structure with no visual artifacts when the hand poses change.

3.2. Hierachical Graph Convolution Network

Our GCN network utilizes a multiresolution graph architecture that follows the subdivision process in Section Sec. 3.1. Different from the single graph GCNs in previous works [20, 25], our GCN network uses different graphs in different layers. At each level, each vertex of the graph corresponds to a vertex on the mesh and the graph topology is defined by the mesh edges. Between two adjunct resolution levels, the network uses the $\mathbf{L_s}$ in Eq. (3) for upsampling operation.

This architecture is designed for scalable inference. When the computing resources are limited, only the low-resolution mesh needs to be calculated; when the computing resources are sufficient, then we can calculate all the way to the high-resolution mesh. Moreover, this architecture allows us to explicitly supervise the intermediate results, so the details would be added level-by-level.

3.3. Graph Frequency Decomposition

In order to supervise the output mesh in the frequency domain and design the frequency-based metric, we need to do frequency decomposition on mesh shapes. Here, we regard the mesh as an undirected graph, and 3D locations of mesh vertices as signals on the graph. Then, the frequency decomposition of the mesh is the spectrum analysis of this graph signal. Following [10], given an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with a vertices set of $\mathcal{V} = \{1, 2, ..., N\}$ and a set of edges $\mathcal{E} = \{(i, j)\}_{i,j\in\mathcal{V}}$, the Laplacian matrix is defined as $\mathbf{L} := \mathbf{D} - \mathbf{A}$, where $\mathbf{A}$ is the $N \times N$ adjacency matrix with entries defined as edge weights a_{ij} and $\mathbf{D}$ is the diagonal degree matrix. The ith diagonal entry $di = \sum_j a_{ij}$.

In this paper, the edge weights are defined as

$$a_{ij} := \begin{cases} 1, & (i, j) \in \mathcal{E} \\ 0, & otherwise \end{cases} \tag{5}$$

which means all edges have the same weights. We decompose $\mathbf{L}$ using spectrum decomposition:

$$\mathbf{L} = \mathbf{U}^\top \mathbf{\Lambda} \mathbf{U}. \tag{6}$$

Here, $\mathbf{\Lambda}$ is a diagonal matrix, in which the diagonal entries are the eigenvalues of $\mathbf{L}$. $\mathbf{U}$ is the eigenvector set of $\mathbf{L}$. Since the Laplacian matrix $\mathbf{L}$ describes the fluctuation of the graph signal, its eigenvalues show how "frequent" the fluctuations are in each eigenvector direction. Thus, the eigenvectors of larger eigenvalues are defined as higher frequency bases, and the eigenvectors of smaller eigenvalues are defined as lower frequency bases. Since the column vectors of $\mathbf{U}$ is a set of orthonormal basis of the graph space, following [34], we define transform $F(x) = \mathbf{U}^\top x$ to be the Fourier transform of graph signal, and $F'(x) = \mathbf{U}x$ to be reverse Fourier transform. This means, given any graph function $x \in \mathbb{R}^{N\times d}$, we can decompose x in N different frequency components:

$$x = \sum_{i=1}^{N} \mathbf{U_i}(\mathbf{U_i}^\top x), \tag{7}$$

where $\mathbf{U_i} \in \mathbb{R}^{N\times 1}$ is the ith column vector of $\mathbf{U}$. d is the dimension of the graph signal on each vertex. $\mathbf{U_i}^\top x$ is the frequency component of x on the ith frequency base.

Having Eq. (7), we can decompose a hand mesh into frequency components. Fig. 1 shows an example of a groundtruth mesh and its frequency decomposition result. The x-axis is the frequencies from low to high. The y-axis is the amplitude of each component in the logarithm. It is easy to observe that the signal amplitude generally decreases as the frequency increases. Fig. 4 shows the cumulative frequency components starting from frequency 0. We can see how the mesh shape changes when we gradually add higher frequency signals to the hand mesh. In general, the hand details increase as higher frequency signals are gradually included.

3.4. Frequency Decomposition Loss

Frequency decomposition loss. Conventional joint and vertex loss, such as the widely used pre-joint error loss [2, 4, 8, 14, 15, 33, 44, 47, 52] and mesh pre-vertex error loss [19, 30, 36, 45] commonly used in human body reconstruction, and Chamfer Distance Loss [1, 17, 26, 42] commonly used in object reconstruction and 3D point cloud estimation, all measure the error in the spatial domain. In that case, the signals of different frequency components are aliased together. As shown in Fig. 1, the amplitudes of

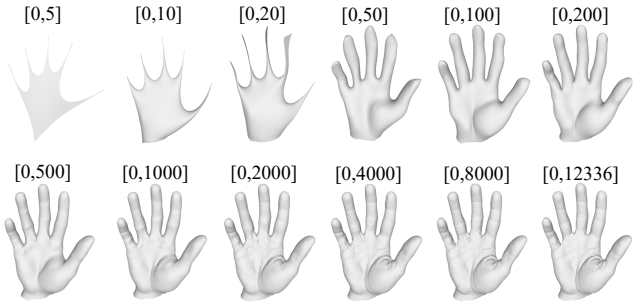

Figure 4. Frequency decomposition of 3D hand mesh. Cumulative frequency components start from frequency 0. The range shows the frequency band. For example, [0,20] means the signal of the first 21 frequencies (lowest 21) added together. We can see how the mesh shape changes when we gradually add higher frequency signals to the hand mesh. In general, the hand details increase as higher frequency signals are included.

low-frequency signals of hand shape are much larger than high-frequency signals, so when alias happens, the high-frequency signals will get overwhelmed, which means direct supervision on the spatial domain would mainly focus on low-frequency signals. Thus, spatial loss mostly does not drive the network to generate high-frequency details. Our experiments in Sec. 4.4 also demonstrate this.

To generate detailed information without being overwhelmed by low-frequency signals, we designed a loss function in the frequency domain. Specifically, we use graph frequency decomposition (Sec. 3.3) to define our frequency decomposition loss as

$$L_F = \frac{1}{F}\sum_{f=1}^{F} \log\left(\frac{\left\|\mathbf{U}_f^\top \hat{V} - \mathbf{U}_f^\top V_{gt}\right\|^2}{\left\|\mathbf{U}_f^\top \hat{V}\right\| \left\|\mathbf{U}_f^\top V_{gt}\right\| + \epsilon} + 1\right), \quad (8)$$

where $F = N$ is the number of total frequency components, $\mathbf{U}_f$ is the fth frequency base, $\|\cdot\|$ is L2 norm, $\epsilon = 1 \times 10^{-8}$ is a small number to avoid division-by-zero, $\hat{V} \in \mathbb{R}^{N\times 3}$ and $V_{gt} \in \mathbb{R}^{N\times 3}$ are the predicted and groundtruth vertex locations, respectively. During training, for every frequency component, our loss reduces the influence of the amplitude of each frequency component, so that information on different frequency components would have equivalent attention. In Sec. 4.3, we demonstrate the effectiveness of the frequency decomposition loss.

Total loss function. We define the total loss function as:

$$L = \lambda_J L_J + \sum_{l=1}^{3}\left[\lambda_v^{(l)} L_v^{(l)} + \lambda_F^{(l)} L_F^{(l)}\right], \quad (9)$$

where l is the resolution level. $l = 1$ is the lowest-resolution level and $l = 3$ is the highest resolution level. $L_J^{(l)}$ is 3D joint location error, $L_v^{(l)}$ is per vertex error, and $L_F^{(l)}$ is the frequency decomposition loss. $\lambda_J^{(l)}$, $\lambda_v^{(l)}$, and $\lambda_F^{(l)}$ are hyper-parameters. For simplicity, we refer $L_J^{(l)}$, $L_v^{(l)}$, and $L_F^{(l)}$ as L_J, L_v, and L_F for the rest of the paper.

Following previous work [9, 45], we define 3D joint location error and per vertex loss as:

$$L_J = \frac{1}{N_J}\sum_{j=1}^{N_J} \|\hat{J}_j - J_{gt,j}\|, L_v = \frac{1}{N}\sum_{i=1}^{N} \|\hat{v}_i - v_{gt,i}\|, \quad (10)$$

where $\hat{J}_j$ and $J_{gt,j}$ are the output joint location and groundtruth joint location. N_J is the number of joints. $\hat{v}_i$ and $v_{gt,i}$ are the estimated and groundtruth location of the ith vertex, and N is the number of vertices.

4. Experiments

4.1. Datasets

Our task requires detailed hand meshes for supervision. Because of the difficulty of acquiring 3D scan data, this supervision is expensive and hard to obtain in a large scale. One alternative plan is to generate meshes from multiview RGB images using multiview stereo methods. Considering the easy access, we stick to this plan and use the generated mesh as groundtruth in our experiments. We do all our experiments on the InterHand2.6M dataset [29], which is a dataset consisting of multiview images, rich poses, and human hand pose annotations. The dataset typically provides 40-100 views for every frame of a hand video. Such a large amount of multiview information would help with more accurate mesh annotation. Finally, we remesh the result hand mesh into the same topology with our 3-level hand mesh template, respectively, so that we can provide mesh supervision for all 3 levels of our network. We use the resulting mesh as groundtruth for training and testing. In this paper, we use the mesh results provided in [28], which are generated using multiview methods of [12], and only use a subset of InterHand2.6m, due to the large number of data in the original dataset. The remeshing method and more dataset details can be found in supplementary material Section IV. In Fig. 8 (last column, "groundtruth"), we show a few examples of the generated groundtruth meshes. Although these meshes are not the exact same as real hands, it is vivid and provides rich and high-fidelity details of human hands. This 3D mesh annotation method is not only enough to support our solution and verify our methods, but is also budget-friendly.

4.2. Implementation Details.

We follow the network architecture in [9] to generate intermediate MANO results. We use EfficientNet [39] as a backbone. The low-level, mid-level, and high-level features are extracted after the 1st, 3rd, and 7th blocks of EfficientNet,

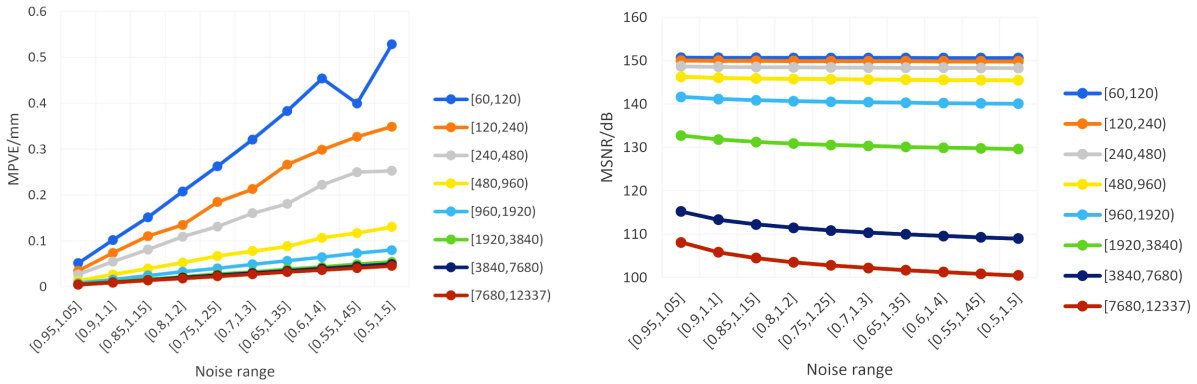

Figure 5. Evaluations using Euclidean distance and MSNR under different noise amplitudes in every frequency band. Each line of different color indicates a frequency band. The maximum and minimum frequencies are shown in the legend. On each line, every dot means adding a random amplitude noise to the mesh. The noise amplitude of each dot is evenly distributed in the ranges shown on the x-axis. The result validates that Euclidean distance is more sensitive to error in low-frequency bands, and MSNR is more sensitive to noise in high-frequency bands. Thus, compared to Euclidean distance, MSNR can better measure the error in high-frequency details.

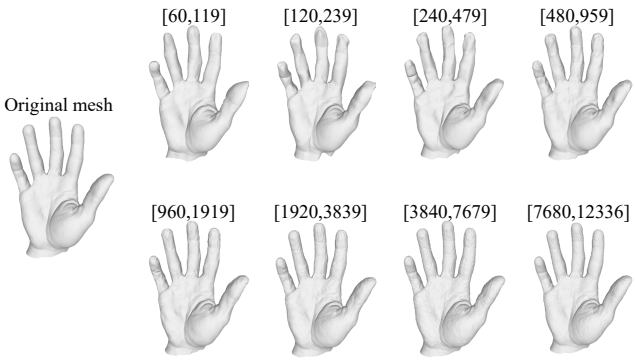

Figure 6. We show examples of Noisy Meshes. The meshes from left to right are meshes with a noise maximum amplitude of 0.6 and the frequency band changed from [60,119] to [7680,12336]. For easier visualization, we visualize the vertices location changes 5 times larger.

respectively. For each image feature, we use 1×1 convolutions to deduce dimensions. The channel numbers of 1×1 convolution are 32, 32, and 64 from low-level to high-level, respectively. After that, we project the initial human hand vertices to the feature maps, and sample a feature vector for every vertex using bilinear interpolation. The GCN graph has 778, 3093, and 12337 vertices at each resolution level.

In the training process, we first train [9] network, and then use the pretrained result to train our scalable network. For training [9], we use their default hyper-parameters, set the learning rate to 1×10^{-4}, and set batch size to 48. When training GCN network, we set λ_J to be 1, set $\lambda_v^{(1)}$ and $\lambda_F^{(1)}$ to be 1 and 60, set $\lambda_v^{(2)}$ and $\lambda_F^{(2)}$ to be also 1 and 60, and set $\lambda_v^{(3)}$ and $\lambda_F^{(3)}$ to be 1 and 100. The learning rate is set to 5×10^{-4} for GCN and 1e-4 for the rest of the network. The batch size is set to 28. The training process takes about 25 hours on 1 NVIDIA GTX3090Ti GPU for 150 epochs. In reference, we use a smooth kernel to post-process the mesh to reduce sharp changes. More details of post-processing will be found in Supplementary materials Section III.

Method	MPJPE/mm ↓	CD/mm ↓	MSNR ↑
MANO	13.41	6.20	-2.64
Ours-level 1	13.25	5.53	-2.70
Ours-level 2	13.25	5.49	-2.62
Ours-level 3	**13.25**	**5.49**	**-0.68**

Table 1. Results on InterHand2.6M dataset. For MPJPE and CD, lower is better. For MSNR, higher is better. As shown in the table, the proposed method improves the accuracy of hand surface details. While our method generates better shape details in a scalable manner, the joint locations and the overall shape also become slightly more accurate.

Level	#parameter	GFLOPS	#vertices	#faces
baseline	14.5M	1.8	778	1538
1	14.5M	1.9	778	1538
2	14.5M	2.5	3093	6152
3	14.7M	4.8	12337	24608

Table 2. The mesh size and the resources needed for generating different resolution levels of meshes.

4.3. Quantitative Evaluation

We use mean per joint position error (MPJPE) and Chamfer distance (CD) to evaluate the hand pose and coarse shape. Besides, to better evaluate personalized details, we also evaluate our mesh results using the proposed mean signal-to-noise ratio (MSNR) metric.

Mean Signal-to-Noise Ratio (MSNR). Previous metrics for 3D hand mesh mostly calculate the Euclidean distance between the results and the groundtruth. Although in most cases, Euclidean distance can roughly indicate the accuracy of the reconstruction results, it is not consistent with human cognitive standards: it is more sensitive to low-frequency errors, but does not perform well in personalized detail distinction or detailed shape similarity description.

Thus, we propose a metric that calculates the signal-to-noise ratio in every frequency base of the graph. We define our Mean Signal-to-Noise Ratio (MSNR) metric as

$$\text{MSNR} = \frac{1}{F}\sum_{f=1}^{F}\log\left(\frac{\left\|\mathbf{U}_f^\top \hat{V}\right\|}{\left\|\mathbf{U}_f^\top \hat{V} - \mathbf{U}_f^\top V_{gt}\right\| + \epsilon}\right), \quad (11)$$

where $F = N$ is the total number of frequency components and S_f is the signal-to-noise ratio of the fth frequency component. $\mathbf{U}_f$, $\hat{V}$, and V_{gt} have the same meaning as in Eq. (8). $\epsilon = 1 \times 10^{-8}$ is a small number to avoid division-by-zero. Thus, the maximum of S_f is 8. By this design, the SNR of different frequency components would not influence each other, so we can better evaluate the high-frequency information compared to the conventional Euclidean Distance.

We designed an experiment on InterHand2.6m to validate the effectiveness of our metric in evaluating high-frequency details. We add errors of 8 different frequency bands to the hand mesh. For each frequency band, the error amplitude is set under 10 different uniform distributions. As shown in Fig. 5, we measure the MPVE and MSNR for every noise distribution on every frequency band, to see how the measured results of the two metrics change with the noise amplitude in each frequency band. The result shows that in the low-frequency part, MPVE increases fast when the noise amplitude increases (the upper lines), but in high-frequency bands, the measured result changes very slowly when the noise amplitude increases. MSNR behaves completely differently from MPVE. It is more sensitive to noise in the high-frequency band than in the low-frequency band. Thus, compared to Euclidean distance, MSNR better measures the error in high-frequency details. Fig. 6 shows a few examples of noisy meshes.

Evaluation on InterHand2.6M dataset. We report mean per joint position error (MPJPE), Chamfer distance (CD), and mean signal-to-noise ratio (MSNR) to evaluate the overall accuracy of reconstructed hand meshes. Tab. 1 shows the comparison among 3 levels of our proposed method and MANO. As shown in the table, the proposed method improves the accuracy of hand surface details by a large margin (as indicated by MSNR). We also observe that, while our method generates better shape details in a scalable manner, the joint locations and the overall shape of the output meshes also become slightly more accurate (as indicated by MPJPE and CD). Here, the MSNR of MANO, Ours-level 1, and Ours-level 2 are calculated after subdividing their meshes into the same resolution as Ours-level 3.

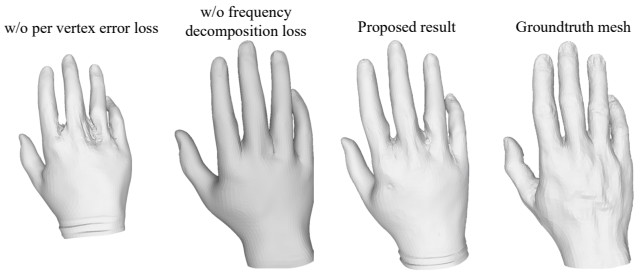

Figure 7. Visualization results of "w/o frequency decomposition loss" and "w/o per vertex error loss" in Sec. 4.4. As shown, if we do not use frequency decomposition loss, the mesh result we get tends to be smoother with less personalized details. If we do not use per-vertex error loss, the mesh's low-frequency information is not well-learned. The mesh we generate will have an overall shape deformation.

Method	MPJPE/mm ↓	CD/mm ↓	MSNR ↑
proposed	**13.25**	**5.49**	**-0.68**
w/o skip connected feature	14.20	5.85	-0.70
w/ average pooling feature	13.95	5.59	-1.10
w/o frequency decomposition loss	14.50	5.86	-1.80
w/o per vertex error loss	14.24	67.8	–0.87

Table 3. Ablation study on the feature skip connection design and the effect of loss functions. From the result, we can see that the frequency decomposition loss helps learn mesh details and the per-vertex error loss helps constrain the overall shape.

4.4. Ablation Study

We conduct several experiments to demonstrate the effectiveness of the feature skip connection design (in Fig. 2). and different loss functions. The results are shown in Sec. 4.3. From the result, we observe that our projection-to-feature-map skip connection design leads to performance improvement in all three metrics. For the loss functions, we observe MSNR degrades when the frequency decomposition loss is removed, indicating inferior mesh details. Removing the per-vertex error loss dramatically increases the Chamfer distance, indicating that the overall shape is not well constrained. The visualization results of the latter 2 experiments are shown in Fig. 7, if we do not use frequency decomposition loss, the mesh result we get tends to be smoother with less personalized details. If we do not use per-vertex error loss, the mesh's low-frequency information is not well-learned. The mesh we generate will have an overall shape deformation.

Scalable design. We also demonstrate the scalable design of the proposed network by analyzing the resource needed at each resolution level (Tab. 2). In general, higher resolution levels require more computational resources in the network,

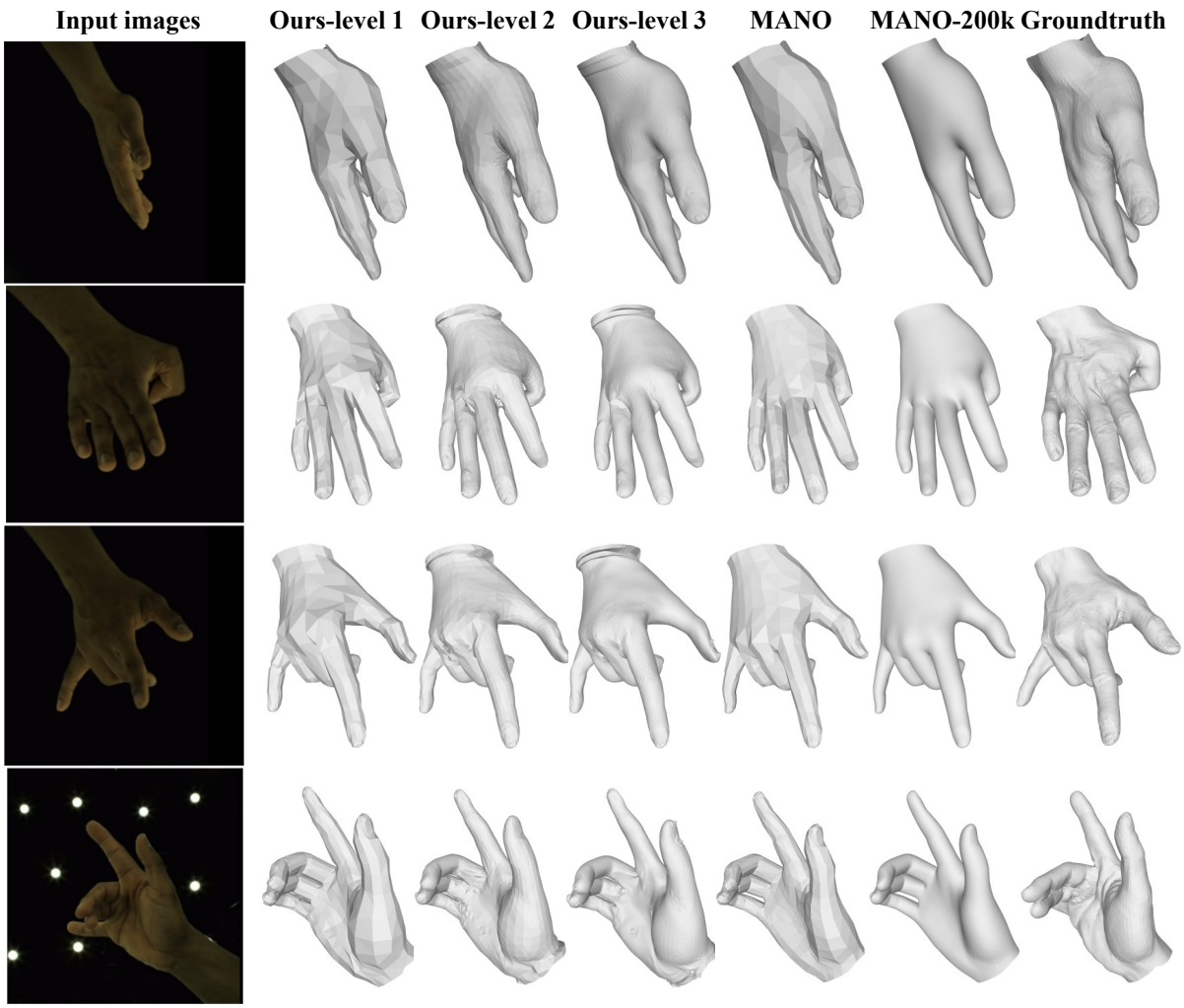

Figure 8. Qualitative reconstruction results. (Best viewed in magnification.) The columns from left to right are input images, our level 1-3 output mesh, MANO mesh, MANO mesh subdivided to 12.3k vertices (same vertex number as our mesh), and groundtruth, respectively. We can see that even if we upsample MANO into the same number of vertices as our mesh, it still does not provide comparable personalized details as our results.

and more resources to store and render the mesh. Still, our approach supports scalable reconstruction and can be applied to scenarios with limited computational resources. Here, "baseline" means only generating the MANO mesh in our network.

Visualization Results. The qualitative reconstruction results are shown in Fig. 8. We observe that even when MANO is upsampled to 200k vertices, it still does not capture personalized details while our results provide better shape details. More qualitative results can be found in the Supplementary Material Section V.

5. Conclusion

We provided a solution to reconstruct high-fidelity hand mesh from monocular RGB inputs in a scalable manner. We represent the hand mesh as a graph and design a scalable frequency split network to generate hand mesh from different frequency bands. To train the network, we propose a frequency decomposition loss to supervise each frequency component. Finally, we introduce a new evaluation metric named Mean Signal-to-Noise Ratio (MSNR) to measure the signal-to-noise ratio of each mesh frequency component, which can better measure the details of 3D shapes. The evaluations on benchmark datasets validate the effectiveness of our proposed method and the evaluation metric in terms of recovering 3D hand shape details.

Acknowledgments

This work is supported in part by a gift grant from OPPO.

References

[1] Vassilis Athitsos and Stan Sclaroff. Estimating 3d hand pose from a cluttered image. In *CVPR*, volume 2, pages II–432. IEEE, 2003. 4

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019. 4

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *CVPR*, pages 6121–6131, 2020. 1

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 4

[5] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013. 2

[6] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. Temporal-aware self-supervised learning for 3d hand pose and mesh estimation in videos. pages 1050–1059, 2021. 1

[7] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. *arXiv preprint arXiv:2112.02753*, 2021. 1

[8] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *CVPR*, pages 10451–10460, 2021. 4

[9] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *CVPR*, 2021. 5, 6

[10] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997. 2, 4

[11] Jian Du, Shanghang Zhang, Guanhang Wu, José MF Moura, and Soummya Kar. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*, 2017. 2

[12] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. 5

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, pages 10833–10842, 2019. 2

[14] Hengkai Guo, Guijin Wang, Xinghao Chen, and Cairong Zhang. Towards good practices for deep 3d hand pose estimation. *arXiv preprint arXiv:1707.07248*, 2017. 4

[15] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, pages 571–580, 2020. 4

[16] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 1, 2

[17] Li Jiang, Shaoshuai Shi, Xiaojuan Qi, and Jiaya Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *ECCV*, pages 802–816, 2018. 4

[18] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 4

[20] Deying Kong, Haoyu Ma, and Xiaohui Xie. Sia-gcn: A spatial information aware graph neural network with 2d convolutions for hand pose estimation. *arXiv preprint arXiv:2009.12473*, 2020. 4

[21] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020. 2

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 1

[23] Charles Loop. Smooth subdivision surfaces based on triangles. *Master's thesis, University of Utah, Department of Mathematics*, 1987. 3

[24] Yi Ma, Jianye Hao, Yaodong Yang, Han Li, Junqi Jin, and Guangyong Chen. Spectral-based graph convolutional network for directed graphs. *arXiv preprint arXiv:1907.08990*, 2019. 2

[25] Jameel Malik, Soshi Shimada, Ahmed Elhayek, Sk Aziz Ali, Christian Theobalt, Vladislav Golyanik, and Didier Stricker. Handvoxnet++: 3d hand shape and pose estimation using voxel-based neural networks. *arXiv preprint arXiv:2107.01205*, 2021. 4

[26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 4

[27] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, pages 752–768. Springer, 2020. 2

[28] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, 2020. 2, 5

[29] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020. 2, 5

[30] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 4

[31] Hao Peng, Chuhua Xian, and Yunbo Zhang. 3d hand mesh reconstruction from a monocular rgb image. *The Visual Computer*, 36(10):2227–2239, 2020. 2

[32] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. 36(6). 1, 2, 3

[33] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *ICCV*, pages 12385–12395, 2021. 4

[34] Stefania Sardellitti, Sergio Barbarossa, and Paolo Di Lorenzo. On the graph fourier transform for directed graphs. *IEEE Journal of Selected Topics in Signal Processing*. 4

[35] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018. 2

[36] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*, 2020. 4

[37] Chao Shang, Qinqing Liu, Qianqian Tong, Jiangwen Sun, Minghu Song, and Jinbo Bi. Multi-view spectral graph convolution with consistent edge attention for molecular modeling. *Neurocomputing*, 445:12–25, 2021. 2

[38] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013. 2

[39] Mingxing Tan and Quoc V Le. Efficientnet: Improving accuracy and efficiency through automl and model scaling. *arXiv preprint arXiv:1905.11946*, 2019. 2, 5

[40] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *ICCV*, pages 11698–11707, 2021. 1, 2

[41] Zhigang Tu, Zhisheng Huang, Yujin Chen, Di Kang, Linchao Bao, Bisheng Yang, and Junsong Yuan. Consistent 3d hand reconstruction in video via self-supervised learning. *arXiv preprint arXiv:2201.09548*, 2022. 2

[42] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. *arXiv preprint arXiv:2111.12702*, 2021. 4

[43] Xiaodan Xing, Qingfeng Li, Hao Wei, Minqing Zhang, Yiqiang Zhan, Xiang Sean Zhou, Zhong Xue, and Feng Shi. Dynamic spectral graph convolution networks with assistant task training for early mci diagnosis. In *MICCAI*, pages 639–646. Springer, 2019. 2

[44] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In *ECCV*, pages 122–139. Springer, 2020. 1, 2, 4

[45] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of human body shape in motion with wide clothing. In *ECCV*, pages 439–454. Springer, 2016. 4, 5

[46] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. *arXiv preprint arXiv:2008.05079*, 2020. 1

[47] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *ICCV*, pages 2335–2343, 2019. 4

[48] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. In *ICCV*, pages 11281–11292, 2021. 2

[49] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. 2

[50] Xinqian Zheng, Boyi Jiang, and Juyong Zhang. Deformation representation based convolutional mesh autoencoder for 3d hand generation. *Neurocomputing*, 444:356–365, 2021. 2

[51] Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *ICLR*, 2020. 2

[52] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017. 4